DLT2: Dating Latin Texts with Deep Learning Techniques

Mitchell Allen,

Paul Hoffmann

University of Colorado Boulder

Linguistic Background

Latin is a classical language that was spoken in Rome. Due to the Roman Empire's expansion, it spread throughout Europe, as well as parts of Asia and Africa. The language was widely employed as the top choice for international communication until the mid-18th century when it was gradually supplanted by English, French, and others. It is still used today by the Catholic Church. Historically, the date when many texts were written is uncertain or completely unknown.

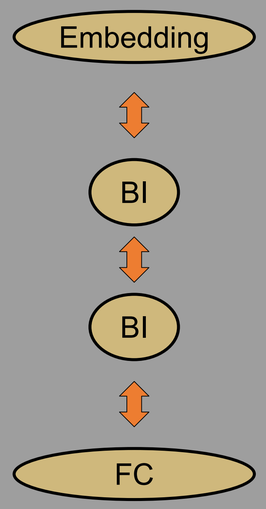
Related Work

A tested approach in a novel context

As far as we are aware, deep learning has not been used to date Latin texts, although there have been several projects attempting to date texts in other languages, including Chinese, Old English, Sanskrit, and Ancient Greek.

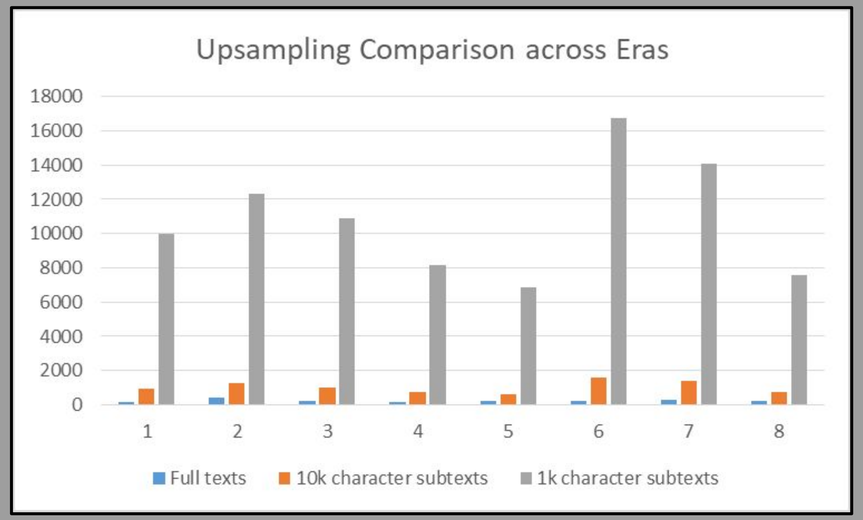
Model Information

- Embedding layer – pre-trained Word2Vec of Latin vocabulary
- Bidirectional LSTM layers – size 1024
- Fully connected layer – dimensions match the number of eras used



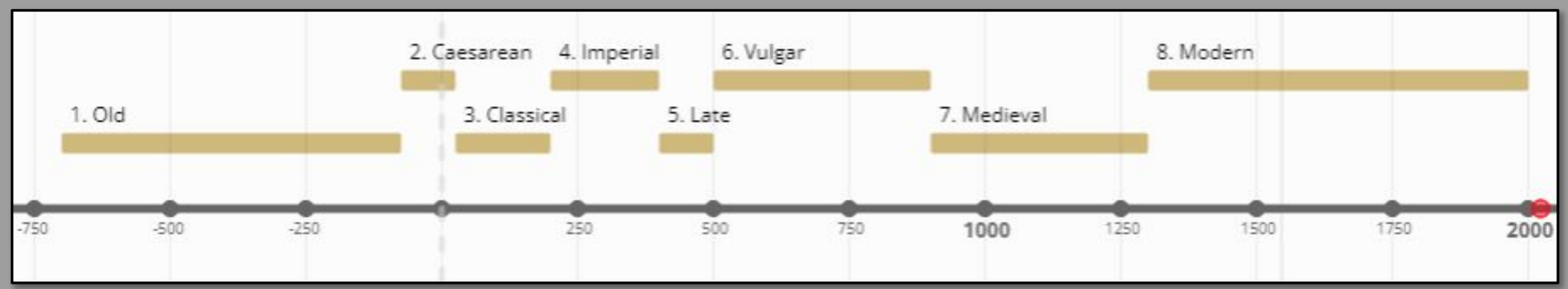
Data

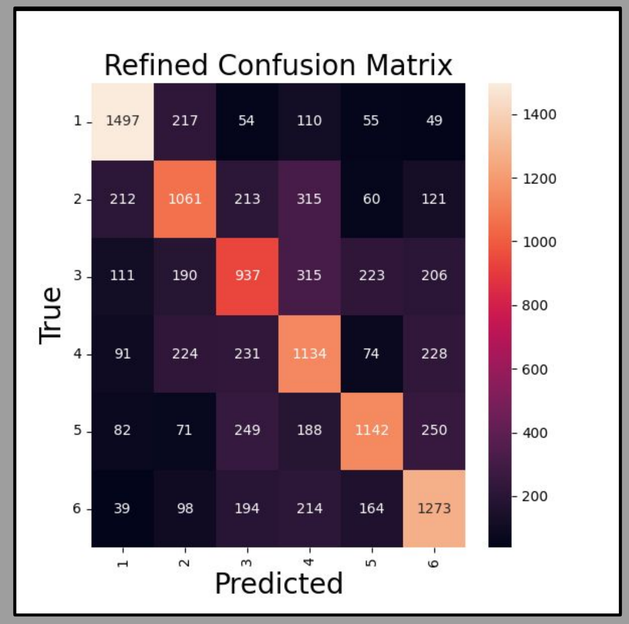
- Model trains on a corpus of 2,220 texts
- Downsampling results in 141 texts in each of 8 eras
- Upsampling is performed through dividing each text into 10,000 character and 1,000 character "subtexts"
- 10k split totals 8,366 subtexts (648 per era)
- 1k split totals 86,577 subtexts (6,885 per era)



Revised Era Distributions

- Eras 2 and 3 combined due to covering the shortest time period and being the most commonly confused
- Modern Latin texts removed due to already being dated and emulation of earlier styles





Confusion matrix from top-performing model

Best Test Results

Accuracy	61%
F1 Score	61%
Precision	63%
Recall	60%

Scan for More

Conclusion

Overall, the model achieved an impressive level of success for this initial foray into applying deep learning to Latin text date approximation. In the future, we would like to create subtext datasets with even smaller subtexts (500 characters) and ask the model to classify texts into narrower time windows. It would also be useful to test the extensibility of this approach to other less-studied classical languages.